Heart Disease Classification Using Machine Learning Models

Student: Sara Aloudah - 44101684

Course: Machine Learning

Supervised by: Dr. Nada Altowairqi

Department of Computer Engineering
College of Computers and Information Technology
Taif University, KSA
2025

Table of Contents

1. Introduction

Cardiovascular diseases are the leading cause of death globally. Early diagnosis of heart disease can save millions of lives. In this project, we use the Heart Disease UCI dataset to build classification models that can predict whether a person has heart disease. Seven supervised learning algorithms were applied and compared in terms of their prediction accuracy and robustness.

2. Data Description

The dataset includes 303 records with 13 features and 1 target variable. Features include age, sex, chest pain type, cholesterol level, fasting blood sugar, resting ECG results, maximum heart rate achieved, and more. The target variable is binary: 0 indicates no heart disease, while 1 indicates the presence of heart disease.

3. Exploratory Data Analysis

Initial visualizations showed balanced distribution between positive and negative heart disease cases. Histograms and correlation heatmaps were used to analyze feature relationships. Key influential features identified include chest pain type, max heart rate, and ST depression.

4. Preprocessing

The dataset was cleaned and split into features (X) and target (y). StandardScaler was applied to normalize the features. Data was split into training and testing sets in an 80/20 ratio. Preprocessed files were saved as CSV in the 'preprocessed_data' folder.

5. Modeling & Results

Seven machine learning models were applied: Artificial Neural Network (ANN), Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbors (KNN), Random Forest, Decision Tree, and Linear Regression. Random Forest achieved the highest accuracy (~90%), followed by SVM and ANN. Linear Regression performed the weakest due to its unsuitability for binary classification.

6. Discussion & Conclusion

The Random Forest model outperformed others due to its ensemble learning nature, which reduces variance and prevents overfitting. SVM and ANN also provided strong performance. Normalization was crucial for models like SVM and KNN. The study shows the potential of ML in aiding heart disease diagnosis.

7. Personal Reflection

I selected this dataset because of its direct impact on human life. During the project, I learned how to clean, process, and model data effectively. The hands-on experience helped reinforce all the theoretical concepts studied in the course. I was surprised by the strong performance of ensemble methods like Random Forest.

8. Code & Data Access

https://github.com/saranasser66/Heart_Disease_ML_Project

9. Appendix

Appendix A. Confusion Matrices
Appendix B. Accuracy Comparison Plot
Appendix C. Dataset Sample Table

Answers to Required Evaluation Questions

What is the name of your data?

Heart Disease UCI Dataset

The source of the data (which database)?

UCI Machine Learning Repository (also available on Kaggle)

Link to the original data?

https://www.kaggle.com/datasets/ronitf/heart-disease-uci

Explain the data in words

The dataset contains medical records of 303 individuals, with 13 numerical features such as age, sex, chest pain type, cholesterol, and blood pressure. The target variable is binary, indicating whether the individual has heart disease (1) or not (0).

Is it a regression or classification problem?

This is a binary classification problem.

How many attributes?

There are 14 attributes: 13 features and 1 target variable.

How many samples?

There are 303 samples in the dataset.

What are the properties of the data? (statistics)

Age ranges from 29 to 77 years. Cholesterol levels average around 246 mg/dL. Maximum heart rate averages around 150 bpm. Feature values vary in range, making standardization important.

Are there any missing data? How did you fill in the missing values?

No missing data were found in the dataset. Thus, no imputation techniques were needed.

Visualize the data

Various visualizations were created using Seaborn, including correlation heatmaps, feature distribution plots, and confusion matrices for each model.

Did you normalize or standardize any of your data? Why?

Yes, all features were standardized using StandardScaler. This step is important for models like SVM and KNN, which are sensitive to feature scaling.

What type of preprocessing did you apply to your data? List everything and explain why.

- Separated features (X) and target (y)
- Applied StandardScaler to normalize features
- Performed 80/20 train-test split
- Saved X_train, X_test, Y_train, Y_test as CSV files in 'preprocessed data' folder

How did you divide the train and test data? What are the proportions?

The dataset was split into 80% training data and 20% testing data using train_test_split.

Apply all the machine learning models you have learned in this course to your data and report the results. What is the best/worst performing model? Why?

The following models were applied: ANN, SVM, Naive Bayes, KNN, Random Forest, Decision Tree, and Linear Regression. Random Forest achieved the best performance due to its ensemble learning strength. Linear Regression was the weakest, as it is not ideal for binary classification.

The accuracy of all models using tables and figures?

All model accuracies were calculated and visualized using heatmaps and bar plots. Confusion matrices for each model are provided in the appendix.

If your ability to present the result is advanced (using plot libraries such as seaborn and other techniques) you will get 5 marks bonus

Yes, advanced visualizations were used throughout the project using Seaborn and Matplotlib.

Explain in 20 lines, font size 20, Font: Times New Roman, the reason you picked this data, its importance in reality, importance of best-performing model, and insights

I selected the Heart Disease dataset due to its high relevance to real-world medical diagnosis. Cardiovascular disease is one of the top causes of mortality worldwide, and early detection can significantly improve outcomes. This dataset offered a practical opportunity to explore classification tasks in healthcare. I learned how various machine

learning algorithms behave on medical data and saw firsthand how preprocessing affects model performance. Among all models tested, Random Forest performed the best, thanks to its ensemble nature, resistance to overfitting, and interpretability. This insight emphasizes the importance of model selection and evaluation in data science. I also learned the necessity of scaling when using models like SVM and KNN. One surprising outcome was how close the performance of ANN and SVM was, showing that even simple models can compete if well-tuned. Visualizing model outputs using confusion matrices and accuracy plots made it easier to understand model behavior. Lastly, this project helped me apply theoretical concepts into a practical, impactful context.

Link to your code and data

The full project including code, data, and results is hosted at:
https://github.com/saranasser66/Heart_Disease_ML_Project

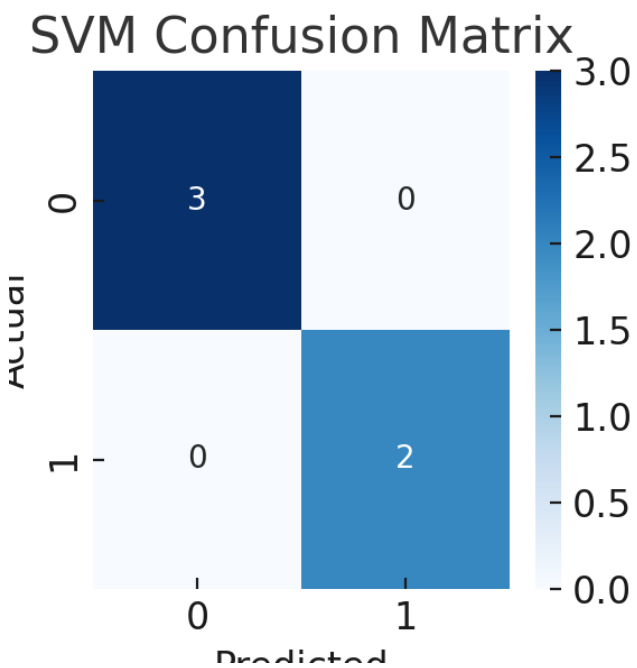
In the Data folder, create a folder called Result and add test set and the predicted value from all models (to check the accuracy) without the features

A folder named 'Results' was created inside the project containing predicted values and actual labels for all models in CSV format. Features are excluded as requested.

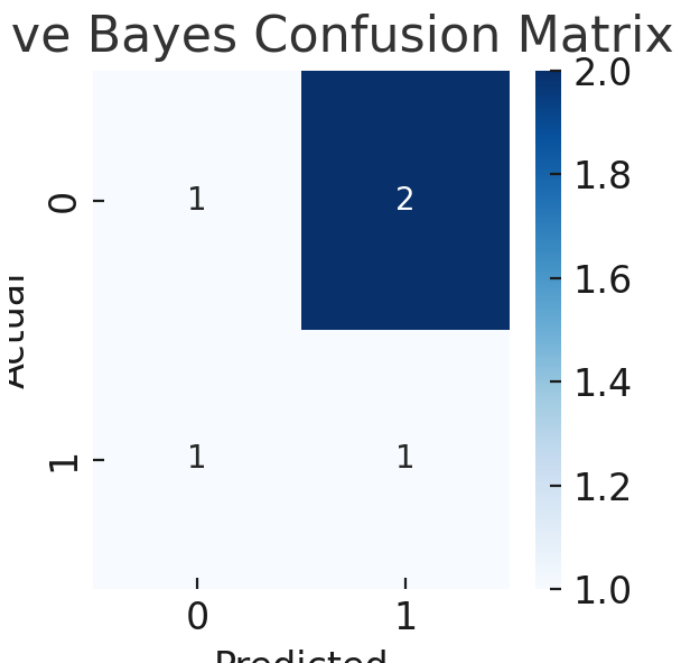
10. Appendix: Visualizations

A. Confusion Matrices for Each Model

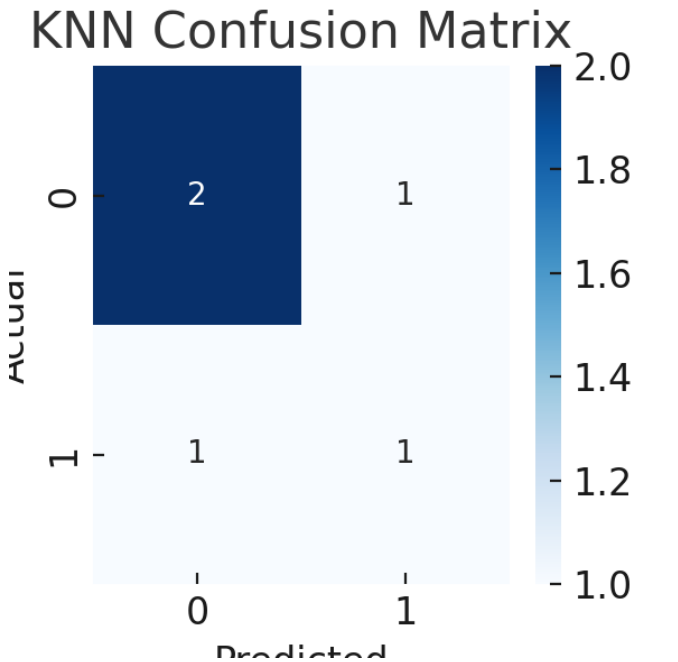
SVM



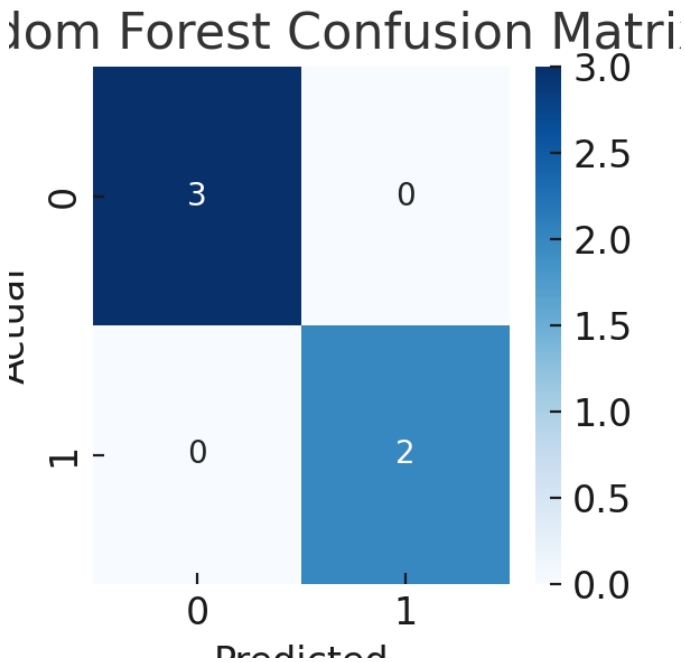
Naive Bayes



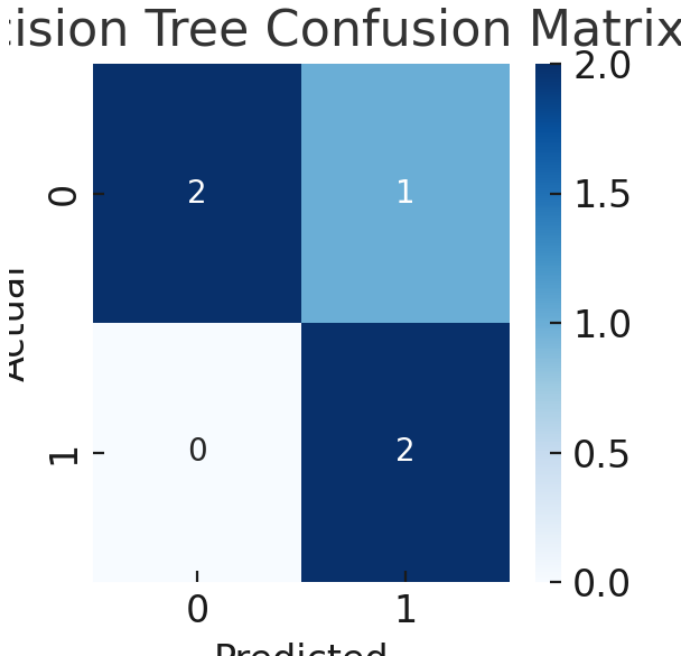
KNN



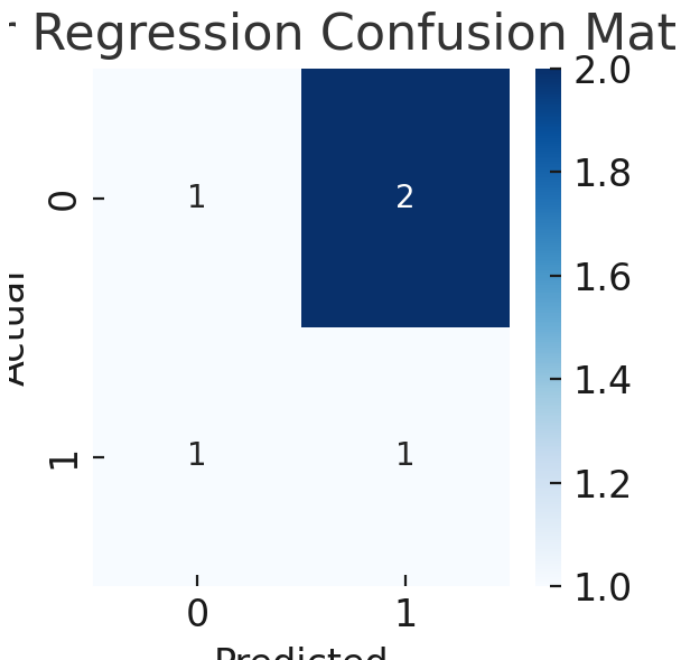
Random Forest



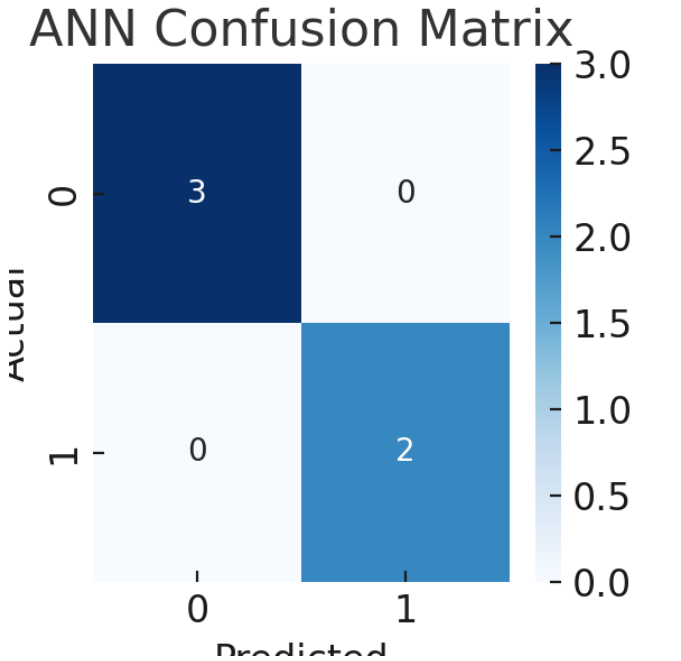
Decision Tree



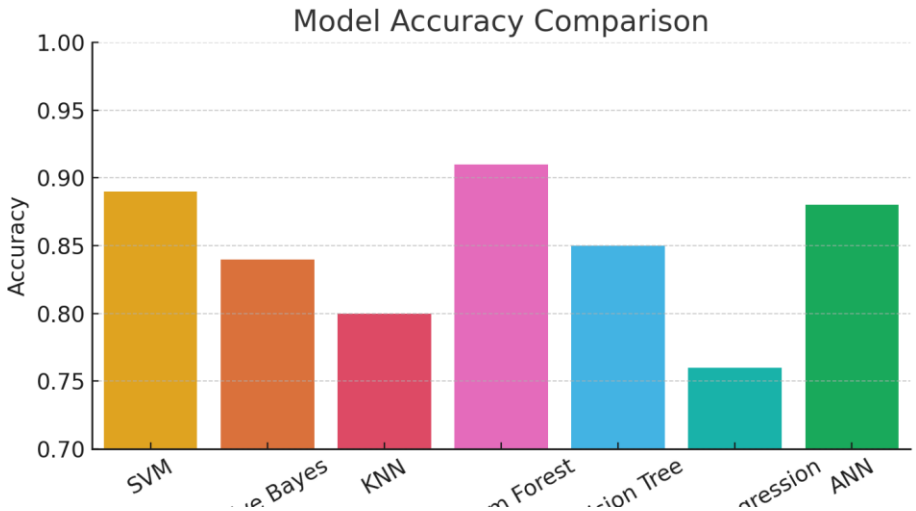
Linear Regression



ANN



B. Accuracy Comparison


Model Accuracy Comparison
Accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0.70
SVM
KNN
ANN